Diary of the Personal Project

Week 1

* **Monday 7-31:** On this day I attended the kick-off. I still don’t know what I will write my project about. I am undecided whether I will take the easy route and repurpose one of my old projects and use the saved time learning R or if I will try to push the limits of what I know. If I do the latter, I would like to code a websitewhere the user can interact in some way to get insights on the data.
* **Tuesday 8/1:** I used today to decide what my project will be about. I browsed through Kaggle’s datasets but nothing sparked my interest. Whilst I generally loath using chatGPT (because it feels lazy), I used the AI to ask me questions and to, then, decide for me what my project should be. This interaction is in the appendix below. In short, the AI suggested I used regression analysis to predict housing prices. This is a funny coincidence because I was introduced to regression analysis with this very analysis: how does a property’s size impact its price.

With this information, I searched through Kaggle to find datasets of the Swedish housing market. I wanted to do my project about Sweden because I’d be more passionate. I finally found a reasonable dataset: **Airbnb prices for Stockholm in 2019**. I can use this dataset to conduct a regression analysis and find out what impacts the pricing of Airbnb properties. I can also showcase descriptive statistics with Tableau. Finally, and to push my limits, I can code a website where the user can input information about their property, to find out what the fair price would be.

* **Wednesday 8-2:** Today I had the meeting with the support group. Aside from giving my input on their projects, Joel recommended my to use the extra time to scrape current Airbnb listings and use that instead of my dataset. With their input, the project has the following objectives:
  1. Run a regression analysis to predict what affects pricing in Airbnb listings. Include a Tableau dashboard with descriptive statistics.
  2. Using Flask, code a website where the user can fill a form with the characteristics of the property they are considering renting out. The website should, then, output the fair price of the property given the results from the regression.
  3. With the extra time, use the Chrome extension Simple Scraper to obtain current Airbnb data. Rerun the regression with that and update the price calculator.
* **Thursday 8-3:** I start working today. The first step, before loading the dataset into a python work environment, was to remove certain unnecessary columns from the dataset with Excel, so as to speed up python’s loading time. Those unnecessary columns were:
  1. Columns with non-categorical text: I don’t know language processing so I cannot analyse these.
  2. Columns that do not contain information about the property. Whilst there may be social drivers of price (for instance, hosts that have high reviews are trusted and can charge more because of this), I am interested only in the characteristics of the apartment being listed.

Secondly, I had to clean the spelling of the Swedish characters Å, Ä and Ö these showed up as other characters.

There was a column for the minimum nights the guests would stay at the accommodation. I turned this into a categorical variable (short\_stay) where if the guests would have to stay for at least 10 days, the value would be zero. One, if otherwise.

I also removed the security fee column but not before checking if there was a correlation between the price and the security fee. It could have been the case where owners charge a bit less in rent, if they could hold a security fee to repair eventual damages done to the property. The corelation coefficient was of -0,032 which is too weak to confirm the aforementioned relationship.

In a similar way, there was a column for the cleaning fee that guests have to pay in addition to the rent. I was interested in knowing if owners charged less rent and used the cleaning fee to cover the rent discount. In reality, the correlation coefficient was 0,247. This may be because bigger properties require more time to clean and are more expensive. Regardless, this column is removed from the regression.

The weekly/monthly prices columns were also removed as the nightly price is more consistent. Some properties offered these prices in a manner similar to discounts where is the guest spent a whole week/month, they would pay less. Likewise, the people included in the nightly rate and extra people fee columns were also removed.

As for a measure of property size, there is indeed the property size expressed in squared meters, however most rows do not include this variable. Instead, I will use either the number of bedrooms and bathrooms or how many people the property can house. I will run the two regressions and compare the results to see which model is more accurate.

* **Friday 8-4:** I mostly did nothing. I had the weekly check out. Most of us are a bit lost.
* **Saturday 8-5:** To make up for not progressing yesterday, I decided to work on a weekend. The objective is to finish the regression to obtain the coefficients. The biggest challenge is the “amenities” column: it includes a set of all the amenities in a listing. I would like to turn it into columns whose values are 1 or 0 if the property offers that amenity. Against my will, I used ChatGPT to help me figure this out. It gave me something that looks like it should work, but for some reason it’s not working with my dataset. I think it’s because of inconsistent data: some amenities are in quotations whilst other aren’t.

After some dialogue with Mr. GPT, I actually managed to get it to work:

1. Parse through the amenities and add them to a set which would, thus, contain the unique amenities.
2. Check if each row contains each unique amenity; creating a new column with the value of one, if yes, or a value of zero if no.

Then, I turn each categorical variable into a dummy variable. I also want to drop one column in each category to avoid the dummy variable trap which would lead to multicollinearity. Then, I run two regressions, one using the “accommodates” column (how many people the apartment can house) and the columns which house the number of beds/bathrooms.

The first regression gave me an R2 of 0,112. Awful. The second one of 0,118… This is because the majority of the amenities are not statistically significant. Using the amount of bedrooms and bathrooms is marginally more accurate so I will use that, I guess.

Week 2

* **Monday 8-7:** The objective today is to begin building the website. Before that, though, I’ll exclude the not-significant variables to see if at least the constant becomes statistically significant.

It didn’t happen… I excluded all non-statistically significant amenities (I kept the dummies for the neighbourhood and the type of housing).

I tried to start the website with Flask. Watched some videos and tried to replicate what they were doing. Nothing worked. My command window looks different. Every tutorial is in IOS.

* **Tuesday 8-8:** Was a bit lazy today. I was planning on not working at all but at the end of the day I got bored. I did some progress and can now build a website that says hi to you.
* **Wednesday 8-9:** Asked ChatGPT for the code to build a calculator where the user inputs a number and the website return that number doubled. Couldn’t make it work….

After lunch I made what ChatGPT gave work. I had to use vscode and the whole thing was a nightmare of errors that I am too stressed to describe. I now have a website that takes a number and doubles it.

After dinner I was bored so I worked on the project again. The website now takes 2 numbers and specifically asks for the amount of bedrooms and bathrooms. It then doubles them and returns the total “price”. Right now the price calculator simply doubles those amounts, but it’ll be easier to do other operations with them now that the “skeleton” is built.

* **Thursday 8-10:** I continued building the website: I added the dropdown menus for where the property is located and what type of housing was being offered (a room, a whole apartment/etc). I also added tickable boxes for the amenities that the listing offers. The challenge now is choosing which ones to include in the website. The thing is: many of the amenities that intuitively should make an apartment more expensive, are actually statistically insignificant. Things like having wifi, allow pets, having a pool, among others do not seem to have a statistically significant effect on the price of the property. At the same time, the host being able to offer “babysitting recommendations” does have a positive impact on the price of the property. I’m intending on adding certain amenities that *sound* like they should have an impact on price but that don’t. They won’t have an effect on the calculator, but the user would miss them as they would think that they would increase the price of the property they’re listing. As such, I’m including three types of amenities:

1. Those that have a statistically significant impact on price (except “babysitter recommendations” since it’s not an amenity, rather it’s a characteristic of the owner)
2. Those that are not statistically significant but sound important for the pricing of a property: wifi, pets allowed, pool, microwave, heating, AC, wheelchair accessible, etc.
3. Amenities that are present on Airbnb listings, but somehow did not make it into the database: sea view and backyard.

The latter two types will not affect price but I think their exclusion would be missed by the user who owns a property that offers those conditions.

Finally, I placed the boxes in a 4x4 grid and added icons that represented them.

It’s all for today. Tomorrow, the plan is to add the numbers to make the calculator fully functional!

* **Friday 8-11:** Today is the day to input the values for the coefficients. I was missing 1 neighbourhood and one housing type. I had to drop these to avoid the dummy variable trap. I can actually drop the intercept instead of excluding one of the dummies. The intercept is not statistically significant anyway. This changed the numbers from the dummies slightly but R2 is still very low.

Week 3

* **Monday 8-14:** I didn’t do anything because I didn’t want to do anything. Apparently, most people haven’t even decided what they wanna do and here I am with a functional albeit ugly website.
* **Tuesday 8-15:** Week 3’s objective is to scrape Airbnb for new and recent data. I will use Chrome’s extension simple scrape recommended by Joel.

The tool seems to only scrape “elements” in a page. I’d have to open the 8k listings to get the details. I watched some videos to find an alternative, but none seemed to help so I gave up for today.

* **Wednesday 8-16:** I found a way to scrape listings from Airbnb: Apify has an API or something that scrapes the website for the info that I need. The output has a bit of a weird format so I’m reformatting the JSON file to fit it neatly into a table. I think I can get all the needed info.

After formatting everything right and running the regression, I got a much higher R2. The problem is that most coefficients (notably bedrooms and bathrooms) are statistically insignificant… The plan, now, is to select only the few amenities that I included in the website; I might be overfitting the model.

I then tried to include only the amenities that are relevant. This time, by including only 20 amenities and all the dummies for where the listing was and what type of housing, R2 is 54% and both bedrooms and bathrooms have statistically significant coefficients. Tomorrow, I’ll update the numbers on the website.

* **Thursday 8-17:** I started by re-running the regression in crowns instead of dollars, however, when running it in the Swedish currency, the coefficients stop making sense even if I only include the relevant amenities. I though this could be because price is in terms of thousands and all X variables are smaller than three thus skewing the results. As such, I tried logging the price to make it more comparable to the explanatory variables in terms of dimension. It didn’t work: bedrooms was still statistically insignificant and extra bathrooms double the price of a listing. I will continue in dollars, then.

Another option to consider is re-running the 2019 regression using only the relevant amenities. 2023 numbers are more actual, but the data is less neat: the API gives me houses outside of Stockholm kommun which clutter the data. Also, n=360.

* **Friday 8-18:**  I tried once again to find a way to scrape Airbnb’s website but I wasn’t very successful. Next week, the plan is to prepare the analysis from the results. Try to ask for help about which data to use (2019/2023) or see if Viktor can help me scrape Airbnb.

Week 4

* **Tuesday 8-22:** Viktor gave me some tips to improve the regression. I have to invest in some descriptive statistics to get to know my set better. I’ll stick to 2019 and look for outliers to remove them.

Week 5

* **Monday 8-28:** Today I re-tried log-transforming the Y variable. R2 jumped to 50% which is good but certain coefficients are too high: an apartment in Södermalm is estimated to increase the price listed by 400%. Likewise, the listing being an entire apartment/house would increase the price by 14135%... The next step is to remove one dummy per category and add an intercept so that we can estimate the price of a property using the regression and not commit the dummy variable trap.

I’m leaving out the category hotel room (as it’s the least interesting) and for the neighbourhoods, I’ll exclude Hässelby-Vällingby.

Rerunning the regression shows that the R2 stayed the same, however, all the coefficients now are smaller. Taking the examples above: a listing in Södermalm now costs 90% more (down from 400%). Likewise, an entire apartment/house (ie. A not shared accommodation) now costs 49%. This is, however, not statistically significant. The next step is to do some descriptive statistics to exclude outliers.

* **En bild som visar text, skärmbild, linje, Graf

  Automatiskt genererad beskrivningTuesday 8-29:** I’ve done a bar graph to see the proportion of the listings that offered the most common amenities. Like I suspected, the overwhelming majority of properties have a kitchen and wifi and that might be why the effect of these amenities is not seen.

The first histogram of the price column indicates that there is a exorbitantly expensive property that may skew the results. Indeed, there is a property that is listed at 80002. This may be a data error or a palace. Regardless, it is an outlier that should be removed. In fact, the 75% quantile is only 1295. There are, in fact, 8 listings that cost more than SEK 10 000. They should be removed as they are skewing the results.

After excluding the outliers, the histogram looked less skewed, but still had a good number of properties that were listed at a price far higher than the median. As such, I will exclude all entries above the 95% quantile.

After running the regression, kitchen, heating, pets allowed and free street paring are still negative. This can be because kitchen and heating are expected on most accommodations. As for pets being allowed, it can be the case that owners of more expensive properties refuse renters with pets to avoid damages.

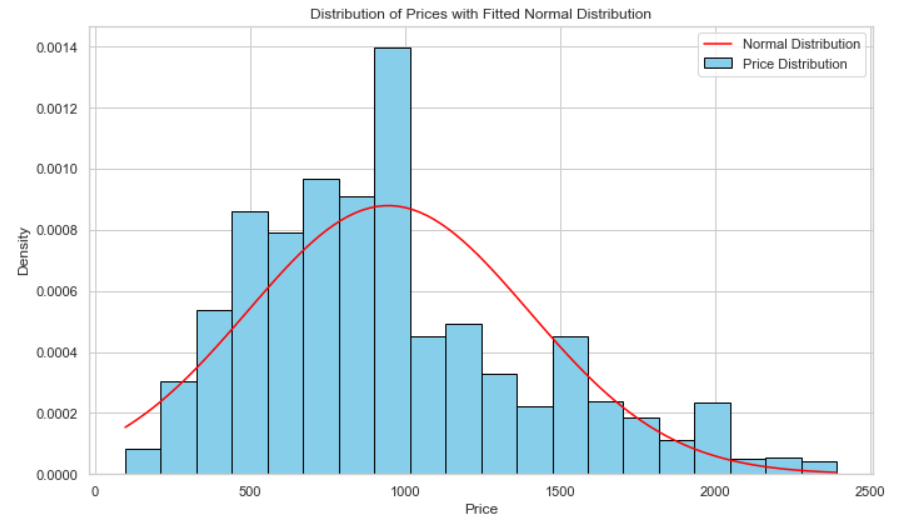
* **Wednesday 8-30:** I plugged in the numbers from the regression onto the website. The page now calculates the price correctly based on the log-lin regression. Prices seem cheap but since the data is from 2019… An option would be to see by how much, on average, prices went up by and multiply the final price by that amount and obtain the final-final price. Otherwise, I could calculate the average price from the 2023 sample, compare it to the 2019 average, and increase the price shown on the website by that amount.

Also, the website now had a (small) image of Stockholm with the divisions. Might need to make it bigger by adding a button to click and zoom, but I’ll save that for tomorrow.

Never mind. I did it today.

Week 6

* **Monday 9-4:** This week I need to have the analysis ready. For that, today I’ll write down what needs to be included in it.
  + Price distribution of the listings
  + Average price per neighbourhood (Use excel to get the colours and paint them on a map)
  + Showcase what impacts the pricing of the listings
  + Show the popularity of the amenities
* **Wednesday 9-6:** Today I went to school to work together with the group on our respective projects.
  + Concerning price distribution of the listings, the prices average at around SEK 1000 per night. There are more listings under the average, however the upper tail is longer which shows that there are still a few expensive listings. In other words, cheaper accommodations are more common, but there are a few quite expensive ones.

Week 7

* **Monday 9-11:** Today I added a new tab to the webpage: it’s the fact sheet/the analysis. I managed to have it as text with a picture next to it whose position alternated from left to right of the text for :sparkle: design :sparkle: purposes. I added more charts to the analysis page. After the charts with the descriptive statistics, I introduce the regression divided into three parts: rooms, housing type and location..
* **Tuesday 9-12:** Went to school instead of working. Did correct some errors though.
* **Wednesday 9-13:** Finished up the analysis section. Next steps are to upload everything to github.

Appendix

* ChatGPT’s help with deciding my topic:

**Me:** Ask me questions so I can figure out what my personal project should be (I'm studying data analysis)

**GPT:** Sure! Let's explore some questions to help you discover a suitable personal project in the field of data analysis:

1. What are your main interests or passions? (e.g., sports, finance, healthcare, social issues, gaming, etc.)
2. Do you have access to any specific datasets or APIs that you find interesting or relevant to your interests?
3. What level of complexity are you comfortable with? Are you looking for a beginner-friendly project or are you ready to tackle more advanced challenges?
4. Are you interested in exploring predictive modeling, statistical analysis, data visualization, or a combination of these?
5. Is there a particular problem you would like to solve using data analysis, or a question you'd like to answer?
6. Are there any real-world issues or challenges you care about that could potentially benefit from data analysis?
7. Are you more interested in structured data (e.g., databases, spreadsheets) or unstructured data (e.g., text, images, audio)?
8. What tools and programming languages are you familiar with or willing to learn for this project? (e.g., Python, R, SQL, Excel, Tableau, etc.)
9. Do you prefer to work independently, or are you open to collaborating with others on a project?
10. Are there any specific areas within data analysis that you want to gain more experience in? (e.g., machine learning, data visualization, data cleaning, time series analysis, natural language processing, etc.)
11. How much time can you allocate to this personal project? Is it a short-term or long-term endeavor?
12. Have you researched any existing data analysis projects or competitions that you find inspiring or intriguing?

Feel free to take your time to answer these questions, and the responses will help guide you towards a data analysis project that aligns with your interests, skills, and goals.

**Me:**

1. I like gaming, but I don't think I wanna do it about gaming. That's too nerdy, you know?
2. I have access to kaggle and the internet
3. 3.I kinda wanna challenge myself this time (I think)
4. Yes, that'd be cool to do
5. Not really. It's not my job to solve problems lmao
6. Yes, but not from some guy's personal project like-
7. No, it HAS to be strunctured data
8. I want to use python, and I'm thinking of building a little interactive website to showcase the data with Flask
9. Independently
10. Machine learning would be cool? Not a must. I think regression analysis is a must, though 11. I have 6 weeks
11. No, fuck that.

Ok so what should I do my project about?

**GPT:** Based on your preferences and constraints, here's a potential data analysis project idea for you:

**Project Idea: Predicting House Prices with Regression Analysis**

**Project Description:** In this project, you will explore a dataset containing information about various houses, including features such as square footage, number of bedrooms, bathrooms, location, and other relevant attributes. Your goal will be to build a regression model to predict house prices based on these features.

**Steps to Complete the Project:**

1. **Data Exploration and Cleaning:** Start by loading and exploring the dataset to understand its structure and features. Perform data cleaning, handling missing values, and addressing any inconsistencies in the data.
2. **Data Visualization:** Use data visualization libraries like Matplotlib or Seaborn to create informative plots to understand the relationships between different features and the target variable (house prices). This step will help you gain insights into the data and identify potential patterns.
3. **Feature Engineering:** Depending on the dataset, you may need to perform feature engineering to create new features or transform existing ones to improve the performance of your regression model.
4. **Regression Model:** Build a regression model using Python's scikit-learn library. Experiment with various regression algorithms such as Linear Regression, Random Forest Regression, or Gradient Boosting Regression. Compare their performance and choose the one that gives the best results.
5. **Model Evaluation:** Split your dataset into training and testing sets to evaluate the performance of your regression model. Use appropriate evaluation metrics like Mean Squared Error (MSE) or R-squared to assess the model's accuracy.
6. **Building an Interactive Website:** Once you have a well-performing regression model, you can create a simple interactive website using Flask. The website can allow users to input values for different house features and get a predicted house price as output.
7. **Deployment:** Deploy your interactive website to showcase your data analysis project. You can use platforms like Heroku or PythonAnywhere for easy deployment.

**Benefits of this Project:**

* You will get hands-on experience with data analysis, data cleaning, and data visualization.
* You'll learn how to build regression models using Python's scikit-learn library.
* By creating an interactive website, you'll gain experience in web development with Flask, making your project more engaging and presentable.

Remember to take the time to thoroughly document your work, including code comments, explanations of your approach, and visualizations. This project will not only demonstrate your skills in data analysis and machine learning but also your ability to showcase your work in a practical and interactive manner. Good luck with your project!